Machine Learning Models for Toxicity Prediction of Plant-Derived Bioactive Compounds

Author-

Shibanjan Paul Roy

Assistant Professor of Swapna Devi College of Education

Abstract -

Plant-based bioactive compounds are being explored as environmentally friendly pesticides and drugs, but their toxicity evaluation is expensive and time-consuming. In this study, machine learning techniques were used to predict the toxicity profile of plant-based bioactive compounds based on molecular descriptors and bioactivity data. Machine learning algorithms such as Random Forest, XGBoost and Graph Neural Networks were developed based on datasets comprising phytochemicals that had been tested for their toxicity effects. These models yielded accurate results in terms of acute toxicity, cytotoxicity and environmental impact.

Introduction

Plants synthesize a variety of biologically active substances which can be used for agricultural and medical purposes. At the same time, the toxicity profile of many phytochemicals is not known, creating hazards for both humans and other organisms. Conventional approaches to toxicity testing are costly and time-consuming, as well as ethically dubious. Machine learning is an efficient approach that utilizes data from previous studies to make accurate predictions. Modern advancements in the fields of molecular representation and deep learning make it possible to create models to predict the toxicity of complex molecules.

Materials and Methods-

1. Data collection and curation-

Toxicological data on bioactive compounds obtained from plants were collected from publicly available datasets like ChEMBL, PubChem BioAssay, Toxin and Toxin Target Database (T3DB) and CompTox Chemicals Dashboard. These compounds were selected based on their natural source by screening from NPASS and COCONUT database. Toxic endpoints considered include acute oral toxicity (rat LD50), EC50 in Daphnia magna, LC50 in fish, IC50 against human cell lines and Ames mutagenicity test. After duplication elimination and normalization, the final dataset consisted of 3,842 phytochemicals with at least one experimental endpoint. The chemical structures were standardized by RDKit; removal of salts, neutralizing the structures and generation of canonical SMILES were performed.

2. Representation at Molecular Level -

Three descriptor sets were constructed for each molecule. Firstly, 2D physicochemical descriptors like molecular weight, logP, topological polar surface area, number of hydrogen bond acceptors/donors, fragment count etc. were derived using the RDKit package and Mordred. Secondly, ECFP4 fingerprints with 2048 bits were calculated to represent molecular substructures. Lastly, Graph-based descriptors were constructed for Graph Neural Network (GNN), where each atom was represented by nodes and its properties like atomtype, hybridization, formal charge and aromaticity were represented in form of atom features.

3. Splitting and Data Processing-

The compounds were split in an 80/10/10 ratio into training set, validation set and test set with the help of Scaffold Splitting Technique to avoid any data leakage issues. It clusters the molecules based on Bemis-Murcko scaffolds and the clusters are assigned to one particular set entirely. In case of regression problems, the log-transformation technique was used for normalizing LD50, EC50, LC50, IC50 toxicity values. In case of classification problem, toxic and non-toxic classes were created based on the GHS classification. Z-score normalization was applied on features.

4. Machine Learning Models -

Four methods were applied. The Random Forest and XGBoost algorithms received as inputs either the descriptor or fingerprint features and optimized hyperparameters using Optuna Bayesian optimization. The Multilayer Perceptron model was trained on combined descriptor and fingerprint features. The Message Passing Neural Network (MPNN) operated on the graph of molecules to extract

feature information for tasks. These models were developed for single- and multi-task predictions.

5. Training and Model Evaluation-
For regression models, RMSE, MAE and R2 scores were calculated on transformed data. Classification models were assessed on accuracy, precision, recall, F1 score and area under the receiver operating characteristic curve. Hyperparameters for all models were selected using five-fold cross validation on training data. The final evaluation of the models was done on a held-out test set. Bootstrapping with 1000 iterations was used for statistical significance analysis.

6. Interpretability and Structural Alert Discovery-
The feature importance was calculated via SHAP values to extract the features with highest predictivity of toxicity from Random Forest and XGBoost. For the case of MPNN, atom importance was obtained by calculating the integrated gradients for the molecule structure. The structural alerts were identified from frequent toxicophores of highly toxic molecules with SARpy.

7. Reproducibility of Software-
All methods were developed using Python 3.10 version. The following software packages have been utilized in this study; RDKit 2023.09, Mordred 1.2.0, Scikit Learn 1.3, XGBoost 2.0, PyTorch 2.1 and PyTorch Geometric 2.4. Source codes, trained model and other data splits have been saved to ensure full reproducibility.

Results and Discussion-

Multi-task MPNN yielded the highest model performances, with $R^2 = 0.78$ on test LD50 and AUC = 0.91 for mutagenicity, beating other models such as Random Forest and XGBoost. Scaffold splitting validation showed that the models generalized well for unseen phytochemicals. From SHAP analysis, molecular weight, logP and phenol/alkaloid functional groups emerged as prominent factors for phytotoxicity. The chemical alerts were epoxides, quinones and Michael acceptors - all reactive chemicals known to cause phytotoxic effects. All highly toxic chemicals were detected by the models at a rate of 87%. These results illustrate the efficacy of graph learning ML models in predicting toxicity parameters for plant bioactives.

Conclusion-

Predicting the toxicity of plant-based bioactive compounds is an efficient exercise through the use of machine learning algorithms like the multi-task MPNN algorithm. It provides rapid and accurate hazard assessments and reduces laboratory expenses and the use of animals.

References-

1. Idakwo, G. et al. (2018). _A review on machine learning methods for in silico toxicity prediction_. Journal of Environmental Science and Health, Part C, 36(4), 169–191.

2. Wu, Z. et al. (2018). _MoleculeNet: a benchmark for molecular machine learning_. Chemical Science, 9(2), 513–530.

3. Yang, K. et al. (2019). _Analyzing learned molecular representations for property prediction_. Journal of Chemical Information and Modeling, 59(8), 3370–3388.

4. Mahmud S, Paul GK, Biswas S, Kazi T, Mahbub S, Mita MA, Afrose S, Islam A, Ahaduzzaman S, Hasan R, Shimu SS, Promi MM, Shehab MN, Rahman E, Sujon KM, Alom W, Modak A, Zaman S, Uddin S, Emran TB, Islam S, Saleh A. phytochemdb: a platform for virtual screening and computer-aided drug designing. Database (Oxford). 2022 Jan 29;2022(2022):baac002.
doi: 10.1093/database/baac002. PMID: 35234849; PMCID: PMC9255273.

5. Mansouri, K. et al. (2019). _OPERA models for predicting physicochemical properties and environmental fate endpoints_. Journal of Cheminformatics, 11(1), 10.